Report Part 1: Exploring Player Performance Patterns in the 2024–25 NBA Season

Minu Choi, Kevin Lee, Jason Tung, Kendan Vu

2025-11-03

Introduction

In the game of basketball, there are several factors that affect a player's scoring abilities and playstyle. While points per game (PPG) is a commonly used statistic as a measure, analyzing specific game-by-game factors and extrinsic factors that impact a player's scoring abilities and playstyle can provide deeper insights into a player's overall effectiveness. This project aims to identify key variables that may contribute to a player's scoring capabilities, decision-making, and efficiency. We focus on four main variables: Position, Assists, Turnovers, and Age. These variables capture differing aspects of a player's role and performance. Each player on a team takes on a certain role or position and oftentimes, this influences a player's scoring capabilities. The rationale for this is that the ball can only be in a player's hand for so many times in a single game and certain positions such as point guards (PG) tend to have control of the ball for a majority of the game. Assists and turnovers viewed concurrently highlight playmaking efficiency. Lastly, age may have an effect on a player's playstyle in the sense that a veteran may be focused more on efficient scoring through foul baiting and free throws. With these variables in mind, our research questions are: 1. Do guards, forwards, and centers differ significantly in their average points per game (PPG)? 2. Do players who assist more also turn over the ball more? 3. Do older players rely more on free throws for scoring?

Data Summary

We obtained our data for this project from Basketball Reference, a website run by Sports Reference LLC. Founded in 2004, Sports Reference compiles information from professional sports leagues around the world to "to promote the democratization of sports data." Sports Reference's websites are some of the most popular sports-related websites in the world. A major reason for this popularity is the speed and reliability of the information on Basketball Reference. Since 2018, Basketball Reference has obtained their data from SportRadar, the official statistics provider of the NBA. SportRadar collects data on every NBA game played using a standardized system of cameras set up in every NBA arena. Using this system, they are able to "collect 3D pose tracking data derived from the ball and 29 points on each player's body." These cameras shoot at 60 frames per second and ensure reliable metrics on shots, blocks, assists, advanced stats, and more. Over a whole season, this data is aggregated and posted onto Basketball Reference for sports fans to look at. This means our data set is representative of the population. All games and all players are tracked on Basketball Reference. Finally, our data represents a population, not a sample

We implemented several data cleaning steps to ensure the dataset accurately reflects player statistics. First, we removed the final row as it was a league-wide summary rather than an individual observation, manually inserted by Sports Reference. Next, we dropped rows containing missing values for all variables except the Awards column, since most players do not receive awards and missing values in that field are expected. Finally, players that averaged fewer than 15 minutes per game were filtered out; this was to remove those with insufficient data and focus on representing players with stable statistics. We selected 15 as the threshold based on the summary of that variable. We found that Q1 = 12, Median = 20, and Q3 = 27–28. Overall, this filter removed ~8% of players and left 523 observations which reduced noise while allowing us to balance sample size with statistics reliability. This filter process also explains why we dropped rows with missing values instead of imputing them with mean or median value: all of the players that had missing stats had low playing time, meaning they would have been filtered out regardless. One notable limitation of this dataset is that it only includes statistics from one single NBA season, which may not fully represent players' long-term performance trends or account for some variability across years. Additionally, as we decided to remove players with fewer than 15 minutes per game, we may have excluded developing players or those in smaller roles. This means our results are biased towards starters and key rotation players. In addition, this dataset relies on publicly reported statistics, which are accurate, but do not capture contextual factors such as injuries, team strategies, or specific game situations.

Overall, this dataset is appropriate for answering the research questions because it provides complete, player-level statistics from the latest NBA season. It has the necessary variables that allow for direct comparison of performance metrics. In addition, it includes all player positions and sufficient size, ensuring meaningful variation across groups. These attributes make it appropriate for exploring positional differences, efficiency trends, and relationships between playstyle and performance.

```r
library(readr)
library(ggplot2)
library(dplyr)
```

```r
nbadata <- read_csv("nba_data_cleaned.csv", show_col_type = FALSE)
colnames(nbadata)
```

```
##  [1] "Rk"                 "Player"             "Age"               
##  [4] "Team"               "Pos"                "G"                 
##  [7] "GS"                 "MP"                 "FG"                
## [10] "FGA"                "FG%"                "3P"                
## [13] "3PA"                "3P%"                "2P"                
## [16] "2PA"                "2P%"                "eFG%"              
## [19] "FT"                 "FTA"                "FT%"               
## [22] "ORB"                "DRB"                "TRB"               
## [25] "AST"                "STL"                "BLK"               
## [28] "TOV"                "PF"                 "PTS"               
## [31] "Awards"             "Player-additional"
```

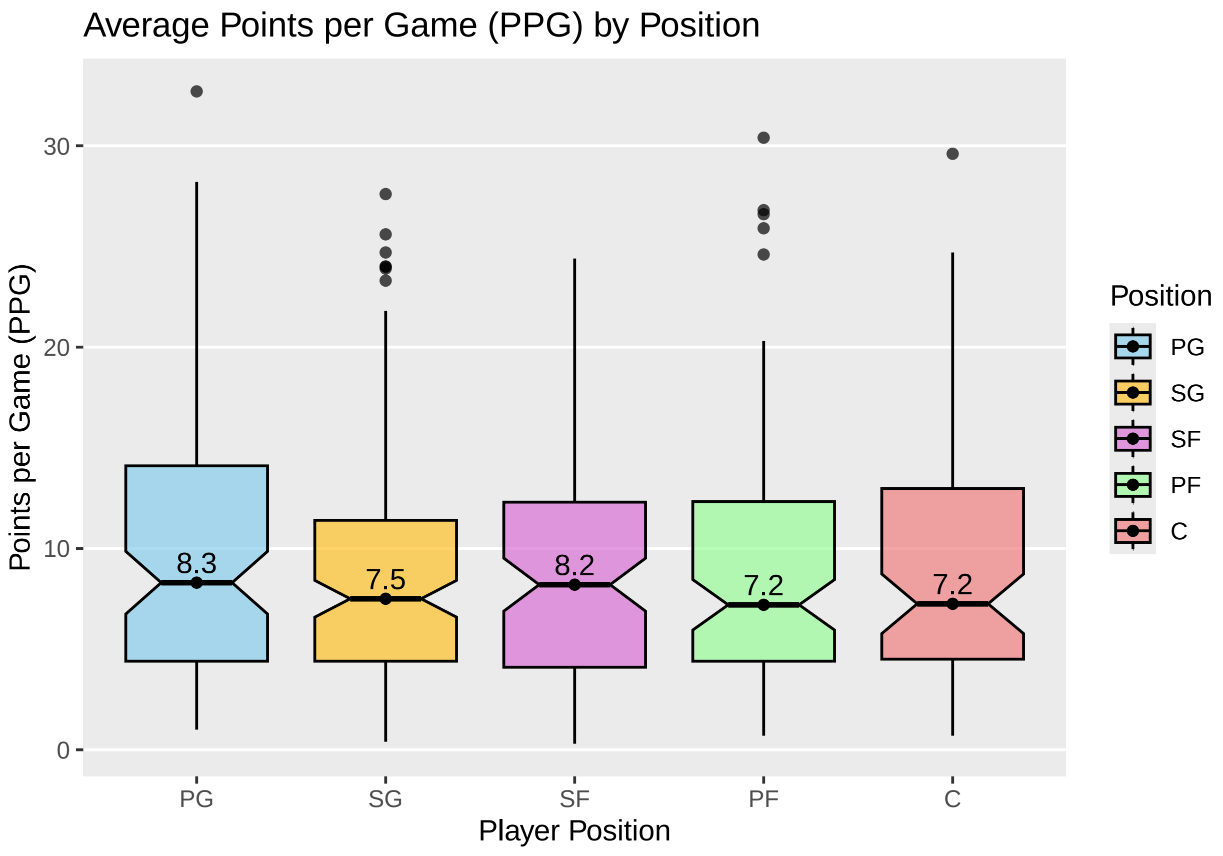
Exploratory Analysis / Conclusion 1

Theme A: Positional Differences in Scoring: Comparing Points per Game across Roles Do guards, forwards, and centers differ significantly in their average points per game (PPG), and how does positional role influence scoring patterns among NBA players?

```r
# Order positions manually to follow the usual order used by the public
nbadata$Pos <- factor(nbadata$Pos, levels = c("PG", "SG", "SF", "PF", "C"))
# Calculate median points per game (PPG) for labeling
median_pos <- nbadata %>% group_by(Pos) %>% summarise(median_PPG = median(PTS, na.rm = TRUE))
by(nbadata$PTS, nbadata$Pos, summary) # 5 number summary for each position
```

```
## nbadata$Pos: PG
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    4.40    8.30   10.28   14.10   32.70 
## ------------------------------------------------------------
## nbadata$Pos: SG
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.400   4.400   7.500   9.076  11.400  27.600
## ------------------------------------------------------------
## nbadata$Pos: SF
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.300   4.100   8.200   9.005  12.300  24.400
## ------------------------------------------------------------
## nbadata$Pos: PF
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   4.400   7.200   9.271  12.325  30.400
## ------------------------------------------------------------
## nbadata$Pos: C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   4.500   7.250   9.262  12.975  29.600
```

```
ggplot(nbadata, aes(x = Pos, y = PTS, fill = Pos)) +
  geom_boxplot(notch=TRUE, alpha = 0.7, color = "black") +
  geom_point(data = median_pos, aes(y = median_PPG)) +
  geom_text(data = median_pos, aes(label = round(median_PPG, 1), y = median_PPG + 1)) +
  scale_fill_manual(values=c("PG"="skyblue","SG"="goldenrod1","SF"="orchid",
                             "PF"="palegreen","C"="lightcoral")) +
  labs(title = "Average Points per Game (PPG) by Position", x = "Player Position",
    y = "Points per Game (PPG)", fill = "Position") +
  theme(legend.position = "right", panel.grid.minor = element_blank(),
        panel.grid.major.x = element_blank())
```

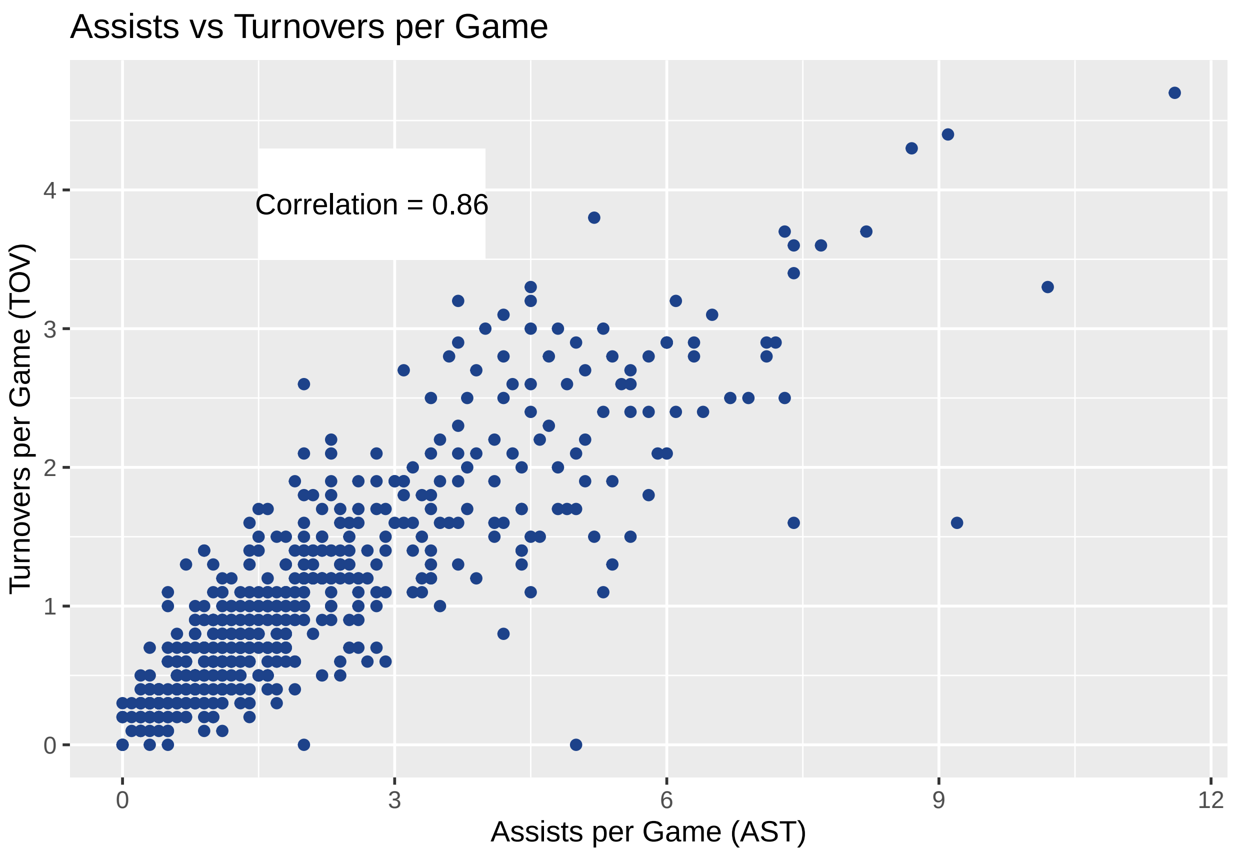


From the summaries and boxplots, point guards have the highest points per the game. With the median for PG being 8.3 PPG, followed by SG at 7.5, SF at 8.2, PF at 7.2, and C at 7.2. The spreads are largest for PG (Q1 = 4.3, Q3 = 14.1) and smaller for SG (Q1 = 4.4, Q3 = 11.4). All positions show long right tails with high-end outliers (PG's max = 32.7, PF's max = 30.4, C's max = 29.6). Additionally, every position has a mean > median, indicating a right-skewed distribution where outliers like a few high scorers can raise the mean. Furthermore, the notched boxplots (used to emphasize median) largely overlap, indicating minimal median differences. In the context of basketball, PGs often tend to control the ball and shoot more, which may explain the higher median as well as its wider variability. PFs and C typically take on lower-volume shots resulting in a lower median, considering they need to fill other roles on the court such as rebounding and screening. Overall, points per game scored by each position seem to differ relatively: PGs tend to score a bit more and have a higher variance due to the volume of shots they take; PFs and C score the least; SFs score in between the two positions mentioned before.

Exploratory Analysis / Conclusion 2

Theme B: Playmaking and Risk: Relationship Between Assists and Turnovers Do players who generate more assists per game also commit more turnovers per game, indicating a trade-off between playmaking activity and possession risk?

```r
cor_value <- cor(nbadata$AST, nbadata$TOV, use = "complete.obs")
ggplot(nbadata, aes(x = AST, y = TOV)) + geom_point(color = "#1D428A") +
  labs(title = "Assists vs Turnovers per Game",
       x = "Assists per Game (AST)", y = "Turnovers per Game (TOV)") +
  annotate("rect", xmin = 1.5, xmax = 4, ymin = 3.5, ymax = 4.3,
           fill = "white", color = "transparent") +
  annotate("text", x = 2.75, y = 3.9, label = paste("Correlation =", round(cor_value, 2)),size = 4)
```



This scatterplot depicts the relationship between assists per game and turnovers per game for players, where each point represents an individual player. The strong positive correlation (r=0.86) indicates that players who generate more assists also tend to commit more turnovers. This pattern, therefore, reflects the nature of playmaking in NBA: players who handle the ball frequently and take greater responsibility in creating scoring opportunities are also more exposed to mistakes and possession risks. Increased offensive involvement comes with both higher productivity and higher turnover risk.

Exploratory Analysis / Conclusion 3

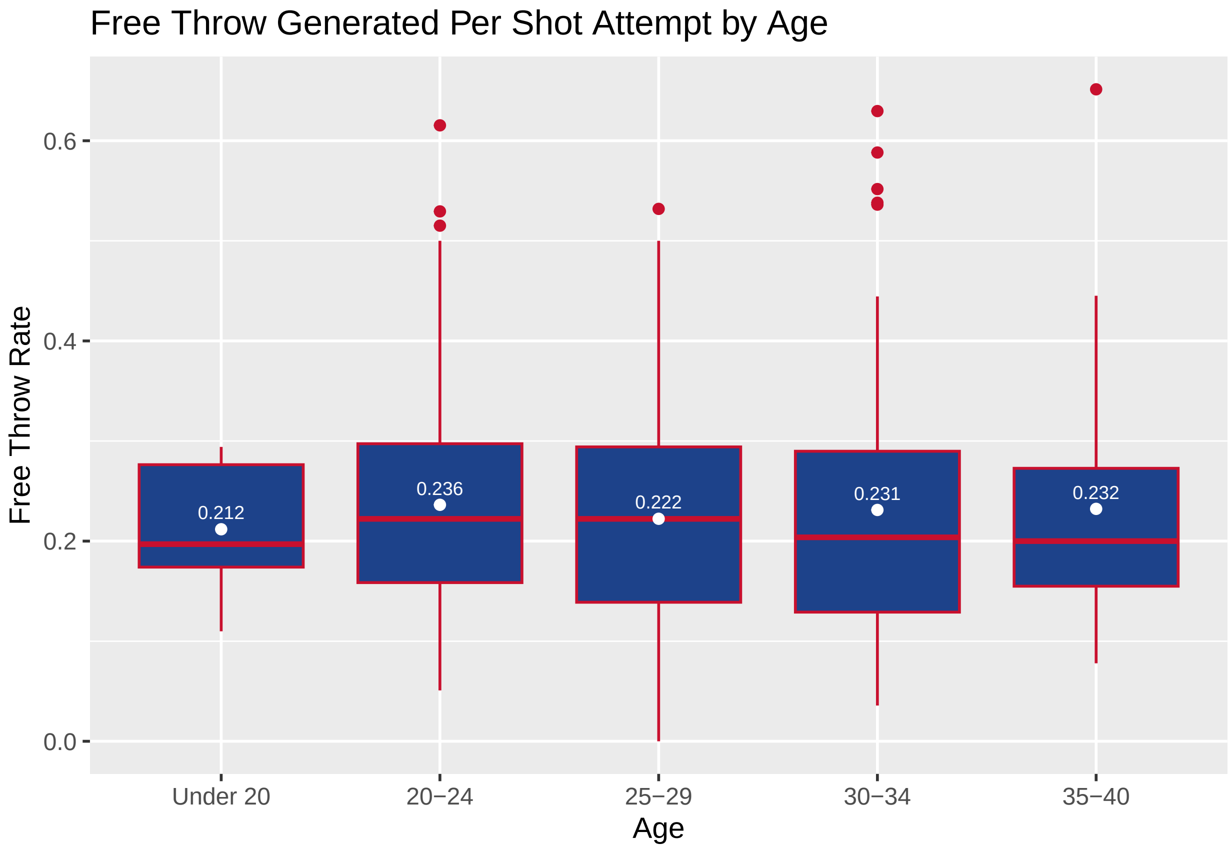
Theme C: Age and Playstyle: Relationship Between Age and Free Throw Reliance Does a player's age significantly affect their free throw rate (the number of free throws generated per shot attempt), indicating a shift in scoring reliance as players get older?

```
#Break the ages into groups
age_breaks = c(0, 20, 25, 30, 35, 41)
#Create names for the groups
age_labels = c("Under 20", "20-24", "25-29", "30-34", "35-40")
#Use cut to split the data into groups
nbadata$Age <- cut(x = nbadata$Age, breaks = age_breaks,
                   labels = age_labels, right = FALSE,include.lowest = TRUE)
```

```
#Add a new column that looks at FT generated off of shot attempts
nbadata$FTRate = nbadata$FTA/nbadata$FGA
#Find mean of FTRate
FTR_mean <- mean(nbadata$FTRate)
```

```
#Find standard deviation of FTRate
FTR_sd <- sd(nbadata$FTRate)
#Calculate the Z-Score for every point
z_scores <- (nbadata$FTRate - FTR_mean) / FTR_sd
#Adds a new column that says if is an outlier in FTRate
nbadata$outliersFTR = abs(z_scores)>3
```

```
#Box plot of Free Throw Rate with outliers removed
age_box_outliersrem = ggplot(filter(nbadata, outliersFTR==F), aes(x=Age, y=FTRate))+
  geom_boxplot(fill="#1D428A", color = "#C8102E")+
  labs(title="Free Throw Generated Per Shot Attempt by Age",
       y="Free Throw Rate")+
  stat_summary(fun.y=mean, geom="point", color = "white")+
  stat_summary(fun.y=mean, geom="text", aes(label=round(after_stat(y), 3)),
               vjust=-0.75, position=position_dodge(0.9), color="white", size = 2.5)
age_box_outliersrem
```



We created this boxplot to understand if age has a significant impact on NBA players' reliance on free throws. To do this, we grouped the players by age on the x-axis. On the y-axis, we plotted free throw rate, which is the number of free throws generated per shot attempt. We also plotted the means using points and labeled the values for a numerical/graphical summary.

We expected that as a player aged they would slow down, with their game become more focused on baiting for fouls. However, we saw that both the median and mean free throw rate across age groups stayed relatively the same. Crucially, the interquartile ranges, represented by the boxes, are of similar size and location, indicating that the variability of free throw rates for the middle 50% of players doesn't substantially change with age. The highest free throw rate was for the 20-24 group, who for every 100 field goal attempts, attempted 23.6 free throws. In contrast, the lowest average free throw rate was for the under 20 group, who attempted only 21.2 free throws per 100 field goal attempts. This indicated to us that NBA players do not rely more on free throws as they age.

Note: Extreme outliers were removed using the 3σ rule, where any data point falling outside $\mu \pm 3\sigma$ was flagged. We did this because we wanted to look at the free throw rate for the average player, not skewed by players who had an extremely small sample size of field goal attempts and free throws.

AI Usage Tool name and version: ChatGPT 5

User / Date of usage: Kevin Lee / Nov 1

Prompt or query: I'm creating a boxplot of Points per Game (PPG) by player position in R using ggplot2. Can you suggest a clean, readable color palette for PG, SG, SF, PF, C?

Output: For color, I recommend: scale_fill_manual(values = c("PG" = "skyblue", "SG" = "goldenrod1", "SF" = "orchid", "PF" = "palegreen", "C" = "lightcoral"))

User / Date of usage: Kendan Vu / Nov 1

Prompt or query: Can you help me find the color codes for the red and blue NBA logo colors for styling and help me fill out the annotate() function to display the correlation value on the top left side?

Output: The official NBA logo colors are: Blue: #1D428A Red: #C8102E. For the annotation, to display the correlation value in the upper-left portion of the plot, you can use: annotate("rect", xmin = 1.5, xmax = 4, ymin = 3.5, ymax = 4.3, fill = "white", color = "transparent") + annotate("text", x = 2.75, y = 3.9, label = paste("Correlation =", round(cor_value, 2)), size = 4)

References

Wagner, James. “From a Church in Philadelphia, Sports Reference Informs the World.” The New York Times, 13 Feb. 2019, www.nytimes.com/2019/02/13/sports/sean-forman-sports-reference.html.

https://www.nytimes.com/2019/02/13/sports/sean-forman-sports-reference.html

“Basketball Reference Now Using Official NBA Data | Sports-Reference.com.” Sports-Reference.com, 2018, www.sports-reference.com/blog/2018/12/basketball-reference-now-using-official-nba-data/.

https://www.sports-reference.com/blog/2018/12/basketball-reference-now-using-official-nba-data/

Schaefer, Rob. “How NBA Teams Are Using Sportradar’s Data Collection Platform This Season.” Sports Business Journal, 27 Oct. 2023, www.sportsbusinessjournal.com/Articles/2023/10/27/nba-sportradar-synergy-sports/.

https://www.sportsbusinessjournal.com/Articles/2023/10/27/nba-sportradar-synergy-sports/

Project Repository URL

https://github.com/UVA-STAT3080/project-nba_dataset (forked from given repo/template, may return 404)

https://github.com/minuuva/stat-3080-project (in case above returns 404; same exact files)